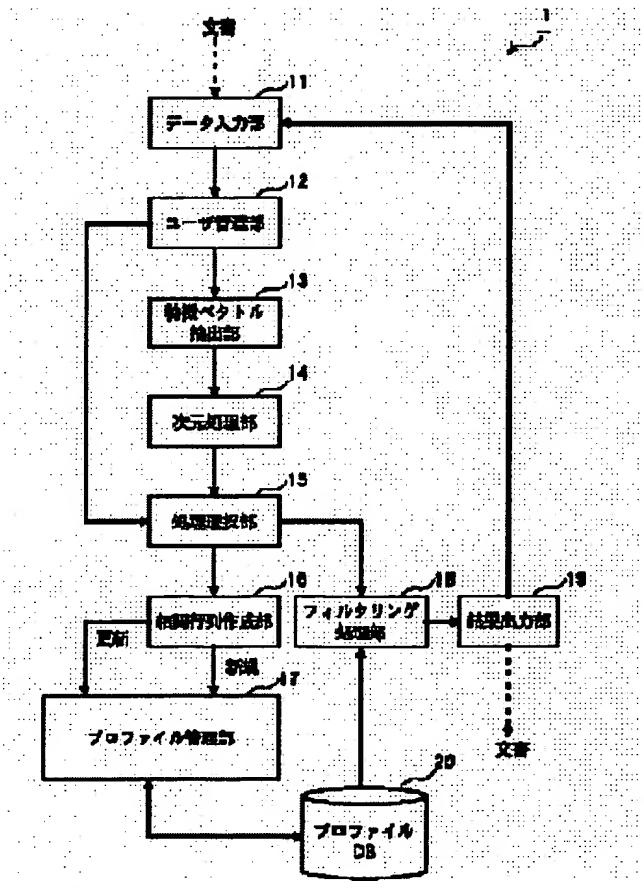
METHOD, DEVICE, AND SYSTEM FOR INFORMATION FILTERING

Patent number: JP11161670
Publication date: 1999-06-18
Inventor: MATSUNAGA TSUTOMU; KIDA HIROMI
Applicant: NTT DATA CORP
Classification:
- international: G06F17/30
- european:
Application number: JP19970329933 19971201
Priority number(s): JP19970329933 19971201

Abstract of **JP11161670**

PROBLEM TO BE SOLVED: To provide the high-precision information filtering device which can automatically reflect a user's interest. SOLUTION: A profile management part 17 generates a user profile from a correlation matrix obtained by dimension deletion from a feature vector set showing features of an input document. A filtering process part 18 calculates the projection of feature vectors by documents and corresponding user profiles and filters the document on the basis of the calculated values. A correlation matrix is generated again from a feature vector set as filtering results and the profile management part 17 updates the corresponding user profiles. Further, the corresponding user profiles are put together to generate a common profile and when a user profile is updated, the corresponding common profile is updated.



(19)日本国特許庁（ＪＰ）　(12) 公開特許公報（Ａ）　(11)特許出願公開番号

特開平11－161670

(43)公開日　平成11年(1999)６月18日

(51)Int.Cl.[6]	識別記号	ＦＩ		
Ｇ０６Ｆ　17/30		Ｇ０６Ｆ	15/403	３４０Ａ
			15/40	３７０Ａ

審査請求　未請求　請求項の数11　ＯＬ　（全 10 頁）

(21)出願番号　特願平9－329933

(22)出願日　平成９年(1997)12月１日

(71)出願人　000102728
株式会社エヌ・ティ・ティ・データ
東京都江東区豊洲三丁目３番３号

(72)発明者　松永　務
東京都江東区豊洲三丁目３番３号　エヌ・ティ・ティ・データ通信株式会社内

(72)発明者　木田　博巳
東京都江東区豊洲三丁目３番３号　エヌ・ティ・ティ・データ通信株式会社内

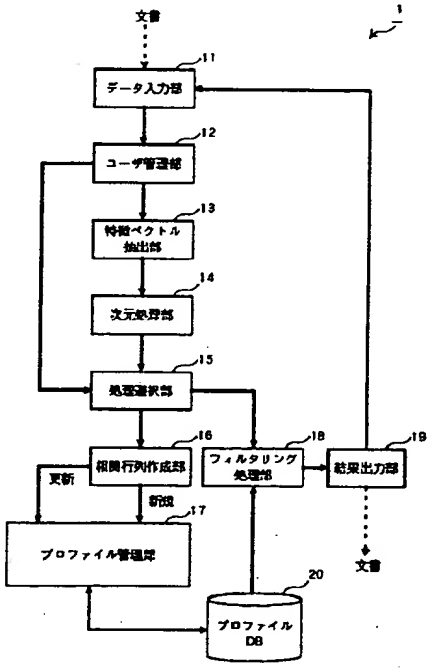
(74)代理人　弁理士　鈴木　正剛

(54)【発明の名称】　情報フィルタリング方法、装置及びシステム

(57)【要約】

【課題】　ユーザの関心が自動的に反映できる、高精度の情報フィルタリング装置を提供する。

【解決手段】　プロファイル管理部１７において、入力文書の特徴を表す特徴ベクトル集合から次元削減して得られる相関行列でユーザプロファイルを作成する。フィルタリング処理部１８は、文書毎の特徴ベクトルと対応するユーザプロファイルとの射影を算出し、算出値に基づいた文書のフィルタリングを行う。フィルタリング結果である特徴ベクトル集合から相関行列を再作成し、プロファイル管理部１７で、対応するユーザプロファイルを更新する。また、対応する各ユーザプロファイルを合成して共有プロファイルを作成し、ユーザプロファイルが更新された場合には、対応する共有プロファイルの更新をも行う。



1

【特許請求の範囲】

【請求項1】　ユーザの関心の有無を識別するための識別情報が付与された電子化情報から冗長な次元を削減した学習ベクトルを抽出し、この学習ベクトルに所定の部分空間類別基準を適用して「関心有」または「関心無」のいずれかのカテゴリに対応するユーザプロファイルを作成する過程と、

選別対象となる新規電子化情報が入力されたときに、その新規電子化情報の特徴を表す対象ベクトルを抽出し、この対象ベクトルと前記作成されたユーザプロファイルとの特徴差を当該ユーザプロファイルに対応する部分空間への射影により求め、この特徴差に基づいて前記新規電子化情報を「関心有」または「関心無」のいずれかのカテゴリに選別する過程と、

選別後の電子化情報から前記ユーザプロファイルと同一形式の更新プロファイルを作成し、この更新プロファイルを用いて前記ユーザプロファイルを更新する過程とを含む、情報フィルタリング方法。

【請求項2】　相互に関連する複数のユーザプロファイルを統合して各ユーザプロファイルと共用関係をなす共用プロファイルを作成する過程をさらに含み、前記選別する過程は、前記共用プロファイルまたは前記ユーザプロファイルとの特徴差に基づいて前記新規電子化情報を「関心有」または「関心無」のいずれかのカテゴリに選別することを特徴とする請求項1記載の情報フィルタリング方法。

【請求項3】　電子化情報の特徴から冗長な次元が削除されたベクトルを抽出するベクトル処理手段と、

ユーザの関心の有無を識別するための識別情報が付与された電子化情報から前記ベクトル処理手段で抽出された学習ベクトルに、所定の部分空間類別基準を適用して「関心有」または「関心無」のいずれかのカテゴリに対応するユーザプロファイルを作成するプロファイル作成手段と、

選別対象となる新規電子化情報から前記ベクトル処理手段で抽出された対象ベクトルと前記ユーザプロファイルとの特徴差を、当該ユーザプロファイルに対応する部分空間への射影により求め、この特徴差に基づいて前記新規電子化情報を「関心有」または「関心無」のいずれかのカテゴリに選別する選別手段とを有し、

この選別手段による選別結果から新たな学習ベクトルを抽出して前記プロファイル作成手段に導くように構成されていることを特徴とする情報フィルタリング装置。

【請求項4】　前記ユーザプロファイルをユーザ毎に管理するユーザ管理手段をさらに備え、このユーザ管理手段が新規ユーザによる最初の選別であることを認識したときに、当該新規ユーザについての初期プロファイル設定用データを対話式で取り込んで前記プロファイル作成手段に当該新規ユーザについての前記ユーザプロファイルを作成させるように構成されていることを特徴とする請求項3記載の情報フィルタリング装置。

2

【請求項5】　前記プロファイル作成手段は、前記抽出された学習ベクトルから部分空間類別基準に基づいて相関行列を作成するように構成されていることを特徴とする請求項3記載の情報フィルタリング装置。

【請求項6】　前記プロファイル作成手段は、前記抽出された学習ベクトルから所定の平均的学習部分空間法の適応的な学習条件に基づいて相関行列を作成するように構成されていることを特徴とする請求項3記載の情報フィルタリング装置。

【請求項7】　前記ベクトル処理手段は、正規直交変換によるKL解析を施して前記冗長な次元を削減するように構成されていることを特徴とする請求項3乃至6のいずれかの項記載の情報フィルタリング装置。

【請求項8】　相互に関わり合う複数の前記ユーザプロファイルを統合して統合前のユーザプロファイルと共用関係をなす共用プロファイルを作成して保存するとともに、この共用プロファイルに関わるユーザプロファイルの少なくとも一つが更新された場合に、当該更新を前記共用プロファイルに反映させる共用プロファイル処理手段を更に備え、

前記選別手段は、前記ユーザプロファイルまたは共用プロファイルを選択的に用いて前記フィルタリングを行うことを特徴とする請求項3乃至7のいずれかの項記載の情報フィルタリング装置。

【請求項9】　前記共用プロファイル処理手段は、統合候補となる複数のユーザプロファイルの各々について関心の有無の差分に着目した距離値を算出し、この距離値の総和から統合するかどうかを判定するように構成されることを特徴とする請求項8記載の情報フィルタリング装置。

【請求項10】　請求項3乃至9のいずれかの項に記載された情報フィルタリング装置を通信回線に接続し、前記通信回線を通じて流通する電子化情報が、前記情報フィルタリング装置に取り込まれるように構成された情報フィルタリングシステム。

【請求項11】　前記情報フィルタリング装置は、エージェント手段を通じて取り込まれた前記電子化情報のフィルタリングを行うように構成されていることを特徴とする請求項9記載の情報フィルタリングシステム。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】本発明は、大量の電子化情報からユーザにとって関心の高いテーマを持つ電子化情報をフィルタリング（フィルタリング）する手法に関する。

【0002】

【従来の技術】近年、インタネットに代表される大規模かつ高速なネットワークの普及等により、エンドユーザが容易に多種の電子化情報を多様な形態で取得できる環

境が提供されている。しかし、情報の電子化の推進は情報化社会の一翼を担う一方、その膨大化した情報は、人間が管理可能な量を遥かに越えてしまう弊害をもたらしており、この問題を解決する手法ないしシステムの開発が望まれている。また、電子化情報の流通化に伴って、大量の電子化情報から必要な情報のみを取捨選択する必要性が生じている。この場合の取捨選択作業は、人手で行うには負担がかかりすぎるため、コンピュータ装置による自動化、例えば、ユーザが関心を持つテーマに沿って、流入する大量の電子化情報を自動的に選別する情報フィルタリング方法に関する検討がなされている。

【０００３】一般に、情報フィルタリング方法では、ユーザの関心度合いを定量化してコンピュータ処理するために、ユーザがどのような情報に関心を有しているかを表現する基準ベクトル（ユーザプロファイルベクトル、ユーザプロファイル、あるいは単にプロファイルとも呼ばれている）が用いられる。ユーザプロファイルは、例えば、予めユーザが関心のある電子化情報に含まれる複数のテキストデータの集合に含まれる単語の出現頻度を単語毎に求め、求めた単語の種類に応じた次元、例えば、単語の種類が１０種類あれば１０次元のベクトルに変換するとともに、これを正規化したものである。

【０００４】また、ベクトルによるパターン認識手法の一形態として部分空間法（部分空間類別法とも呼ばれる）が知られている。この手法は、類別すべきカテゴリを特徴ベクトル成分の分布から形成される部分空間への射影を通して判定する統計的手法である。この場合の変換するベクトル成分の固有ベクトル計算には、例えば、量子化アルゴリズムであるカルーネン・レーベ（Karhunen-Loeve）変換によるＫＬ解析が採用されている。この部分空間法における代表的な手法には、ＣＬＡＦＩＣ(CLASs-Featuring Information Compression)法や、平均学習部分空間法（Averaged Learning Sub-space Method,ＡＬＳＭ）がある。ＣＬＡＦＩＣ法及びＡＬＳＭは、同様の類別基準を持ち、またＡＬＳＭは、対抗するカテゴリも考慮した適応的な学習法である。なお、この部分空間法については、例えば、「パターン認識と部分空間法」（エルッキ・オヤ著、産業図書）等で詳しく記述されている。

【０００５】実際に情報フィルタリングを行う場合は、ユーザが関心有りと判定されるような閾値を予め設定しておき、当該閾値に基づいてユーザプロファイルを参照することにより、対象となる電子化情報群に対して、類似の度合が大きい順にランク付けされる。電子化情報は、例えば、当該ランクの上位から所定数が選択され、ユーザに対して提示される。

【０００６】

【発明が解決しようとする課題】ところで、一般にフィルタリングの誤りには、必要な情報を落とす「漏れ」と、不必要な情報を取り込む「ノイズ」とがあり、これらの間にはトレードオフの関係があることは良く知られたことである。しかし、従来のフィルタリングでは、ユーザの関心度合い、すなわち「必要な情報」のみに着目した一面的なフィルタリングであり、情報の「漏れ」に対する減少のみが考慮されたものである。そのため、「ノイズ」の除去を直接的に考慮しておらず、フィルタリング精度を高める上で限界があった。

【０００７】また、従来のフィルタリングでは、ユーザの関心事項が複数ある場合や、関心の時間的な変化に柔軟に対応することができないという制約があった。具体的には、業務や趣味等はユーザの長期的な関心事項であり、事件等は一時的な関心事項であるが、従来のフィルタリングでは、一様なキーワード入力等によって関心度合いを決定しなければならないために、ユーザの関心の変化に対応した自動的なフィルタリングは不可能であった。

【０００８】そこで本発明の課題は、ユーザの関心情報であるユーザプロファイルを、ユーザ自身が設定、評価する必要なく、変化するユーザの関心に追随する学習機能により、プロファイルの自動作成を可能とし、フィルタリングに係る精度を一定値以上に維持することができる、情報フィルタリング方法を提供することにある。本発明の他の課題は、上記情報フィルタリング方法の実施に適した情報フィルタリング装置を提供することにある。

【０００９】

【課題を解決するための手段】上記課題を解決するため、本発明は、ユーザの関心の有無を識別するための識別情報が付与された電子化情報から冗長な次元を削減した学習ベクトルを抽出し、この学習ベクトルに所定の部分空間類別基準を適用して「関心有」または「関心無」のいずれかのカテゴリに対応するユーザプロファイルを作成する過程と、選別対象となる新規電子化情報が入力されたときに、その新規電子化情報の特徴を表す対象ベクトルを抽出し、この対象ベクトルと前記作成されたユーザプロファイルとの特徴差を当該ユーザプロファイルに対応する部分空間への射影により求め、この特徴差に基づいて前記新規電子化情報を「関心有」または「関心無」のいずれかのカテゴリに選別する過程と、選別後の電子化情報から前記ユーザプロファイルと同一形式の更新プロファイルを作成し、この更新プロファイルを用いて前記ユーザプロファイルを更新する過程とを含む、情報フィルタリング方法を提供する。

【００１０】上記情報フィルタリング方法において、より好ましくは、相互に関連する複数のユーザプロファイルを統合して各ユーザプロファイルと共用関係をなす共用プロファイルを作成する過程をさらに含むようにする。この場合、前記選別する過程は、前記共用プロファイルまたは前記ユーザプロファイルとの特徴差に基づいて前記新規電子化情報を「関心有」または「関心無」の

いずれかのカテゴリに選別する。

【0011】また、上記他の課題を解決する本発明の情報フィルタリング装置は、電子化情報の特徴から冗長な次元が削除されたベクトルを抽出するベクトル処理手段と、ユーザの関心の有無を識別するための識別情報が付与された電子化情報から前記ベクトル処理手段で抽出された学習ベクトルに、所定の部分空間類別基準を適用して「関心有」または「関心無」のいずれかのカテゴリに対応するユーザプロファイルを作成するプロファイル作成手段と、選別対象となる新規電子化情報から前記ベクトル処理手段で抽出された対象ベクトルと前記ユーザプロファイルとの特徴差を、当該ユーザプロファイルに対応する部分空間への射影により求め、この特徴差に基づいて前記新規電子化情報を「関心有」または「関心無」のいずれかのカテゴリに選別する選別手段とを有し、この選別手段による選別結果から新たな学習ベクトルを抽出して前記プロファイル作成手段に導くように構成されていることを特徴とする。

【0012】より好ましくは、前記ユーザプロファイルをユーザ毎に管理するユーザ管理手段をさらに備え、このユーザ管理手段が新規ユーザによる最初の選別であることを認識したときに、当該新規ユーザについての初期プロファイル設定用データを対話式で取り込んで前記プロファイル作成手段に当該新規ユーザについての前記ユーザプロファイルを作成させるように構成する。

【0013】なお、前記プロファイル作成手段は、例えば、前記抽出された学習ベクトルから部分空間類別基準に基づいて、あるいは所定の平均的学習部分空間法の適応的な学習条件に基づいて相関行列を作成するように構成する。前者は、ユーザプロファイルを新規に作成する場合、後者はユーザプロファイルを更新する場合に有効となる。また、前記ベクトル処理手段は、正規直交変換によるKL解析を施して前記冗長な次元を削減するように構成する。

【0014】本発明の他の情報フィルタリング装置は、相互に関わり合う複数の前記ユーザプロファイルを統合して統合前のユーザプロファイルと共用関係をなす共用プロファイルを作成して保存するとともに、この共用プロファイルに関わるユーザプロファイルの少なくとも一つが更新された場合に、当該更新を前記共用プロファイルに反映させる共用プロファイル処理手段を更に備え、前記選別手段が、前記ユーザプロファイルまたは共用プロファイルを選択的に用いて前記選別を行うようにしたものである。なお、前記共用プロファイル処理手段は、例えば統合候補となる複数のユーザプロファイルの各々について関心の有無の差分に着目した距離値を算出し、この距離値の総和から統合するかどうかを判定するようにする。

【0015】上記他の課題を解決する本発明の情報フィルタリングシステムは、上記情報フィルタリング装置を通信回線に接続し、前記通信回線を通じて流通する電子化情報が、前記情報フィルタリング装置に取り込まれるようにしたものである。この場合、前記情報フィルタリング装置は、エージェント手段を通じて取り込まれた前記電子化情報のフィルタリングを行うように構成することが望ましい。

【0016】

【発明の実施の形態】以下、図面を参照して本発明の実施の形態を詳細に説明する。

（第1実施形態）図1は、本発明を適用した情報フィルタリング装置の機能ブロック図である。図中、実線は処理の流れ、破線はデータ類の流れを表す。この情報フィルタリング装置1は、例えばスタンドアロン型のコンピュータ装置の内部または外部に構築される文書データベース（図示省略）及びプロファイルデータベース（以下、プロファイルDB）20、上記コンピュータ装置が所定のプログラムを読み込んで実行することにより形成される、データ入力部11、ユーザ管理部12、特徴ベクトル抽出部13、次元処理部14、処理選択部15、相関行列作成部16、プロファイル管理部17、フィルタリング処理部18、結果出力部19を備えて構成される。また、図示しないが、後述する初期プロファイル用設定データやグループ化基準の設定データ等を対話式に取り込むための設定用インタフェースを搭載した表示装置、文書を取り込むための入力装置、フィルタリング結果を出力するための出力装置をも備えている。

【0017】上記プログラムは、通常、コンピュータ装置の内部記憶装置あるいは外部記憶装置に格納され、随時読み取られて実行されるようになっているが、コンピュータ装置とは分離可能な記録媒体、例えばCD-ROMやFD等に格納された可搬性記録媒体、あるいは構内ネットワークに接続されたプログラムサーバ等に記録され、使用時に読み込まれて上記内部記憶装置または外部記憶装置にインストールされて随時実行に供されるものであってもよい。

【0018】プロファイルDB20は、電子化情報（以下、特にことわらない限り、単数、複数を問わず文書と称する）に対するユーザの関心情報、すなわち「関心有」または「関心無」のカテゴリを表すユーザプロファイル、及び複数のユーザプロファイルと共用関係をなす共用プロファイルをユーザ毎に保存したものである。このユーザプロファイル等は、前述の部分空間法に基づいて作成されるもので、ユーザがアクセスして選別（フィルタリング）を行う度に更新されるようになっている。

【0019】データ入力部11は、ユーザによるアクセスがあったとき、すなわちユーザがスキャナ等の入力装置を通じて文書を本装置に読み込ませたときに、これを取り込んでユーザ管理部12に入力するものである。ユーザ管理部12は、ユーザプロファイル等をユーザ毎に管理しており、上記ユーザからのアクセスを契機に当該

ユーザが予め登録されたユーザか否かを判定し、判定結果に応じて以後の処理を選択的に行う。

【００２０】新規ユーザによるアクセスの場合は、データ入力部１１から入力された文書に初期プロファイル用設定データを割り当てた後、これを特徴ベクトル抽出部１３に入力する。また、新規ユーザからのアクセスである旨を処理選択部１５に通知する。初期プロファイル用設定データは、上記設定用インタフェースを介して対話式で取り込んだ当該新規ユーザの「関心有り」または「関心無し」の識別情報である。一方、既登録ユーザによるアクセスの場合は、入力された文書を特徴ベクトル抽出部１３に入力するとともに、当該ユーザの識別情報を処理選択部１５に通知する。ユーザ管理部１２は、また、プロファイル管理部１７において使用される、後述の共用プロファイル作成のためのグループ化基準の設定をも行う。このグループ化基準は、プロファイルが相互に関連するかどうか、例えばどのユーザとどのユーザの関心が共通するかを表す基準であり、予めシステムパラメータを通じて設定する。上記設定用インタフェースを通じて個々のユーザが明示的に設定するようにすることもできる。

【００２１】特徴ベクトル抽出部１３は、入力された文書の特徴を表すベクトル（特徴ベクトルの集合、またはその集合を代表する一つのベクトル）を抽出する。具体的には、当該文書中に出現するキーワード（以下、単語）の種類を次元数とし、各単語の出現頻度に重みをかけた一または複数のベクトルを演算処理により求める。この場合の単語の重み付けは、公知の「ＴＦ・ＩＤＦ法」により行うことができる。抽出されたベクトルは、次元処理部１４に入力される。

【００２２】次元処理部１４は、特徴ベクトル抽出部１３で抽出されたベクトルに対し、前述のＫＬ解析、すなわち正規直交変換による主成分分析を施し、重みが相対的に低い冗長な次元のベクトルの削減（または次元圧縮とも言う）を行うものである。次元削減されたベクトルは、処理選択部１５に入力される。

【００２３】処理選択部１５は、次元処理部１４で次元削減されたベクトルに対する以後の処理を選択するものである。具体的には、新規ユーザのアクセスの場合にはプロファイル作成、データ入力部１１に残りの入力文書がある場合にはフィルタリング（選別）、フィルタリング完了後の場合はプロファイル更新の処理がそれぞれ選択されるようにする。

【００２４】相関行列作成部１６は、処理選択部１５で選択された処理がプロファイル作成またはプロファイル更新の場合に、次元削減されたベクトルに各々対応した相関行列を作成するものである。

【００２５】プロファイル管理部１７は、主として、プロファイルや共用プロファイルのプロファイルＤＢ２０への保存、読み出し、更新を行うものである。すなわち、相関行列作成部１６で作成された相関行列を当該ユーザの関心を表すユーザプロファイルとして、これをプロファイルＤＢ２２にユーザ毎に保存させ、フィルタリングの際に、このプロファイルをプロファイルＤＢ２０から適宜読み出す。また、プロファイルＤＢ２０中において、ユーザ管理部１２で設定したグループ化基準に対応した複数のユーザプロファイルを統合（該当する複数の相関行列の総和）して当該グループ識別情報の共用プロファイルを作成し、さらに、新規または更新されたユーザプロファイルが、共用プロファイルに関連する場合は、そのユーザプロファイルの情報で対応する共用プロファイルを更新する。共用プロファイルを設けることにより、ユーザの持つ情報を提供し合い、これらのユーザ間で相互に関心情報を共有することが可能となる。つまり、特定のユーザプロファイルに含まれない関心情報があっても、共用プロファイルに当該関心情報が存在すれば、それを補完的に参照できるようになる。

【００２６】フィルタリング処理部１８は、処理選択部１５でフィルタリング処理が選択された場合に、プロファイルＤＢ２２に格納された当該ユーザのユーザプロファイルまたは当該ユーザが属する共有プロファイルによって文書（実際には次元削減されたベクトル）のフィルタリングを行う。フィルタリング結果は、結果出力部１９を通じて出力装置または文書データベースに出力される。結果出力部１９は、また、フィルタリング結果を既存のユーザプロファイルに反映させるためにデータ入力部１１にフィードバックする機能をも有する。

【００２７】次に、本発明によるプロファイル及びフィルタリングの概念を説明する。図２は、プロファイル作成の処理手順図である。この場合のプロファイルは、新規ユーザより入力された文書に基づいて作成されるもので、この文書には、上述のようにして「関心有」または「関心無」の識別情報が付与されている。なお、新規ユーザという場合、ユーザ管理部１２で既に管理されているユーザであるが、そのユーザからのアクセスが全くなく、フィルタリングを初めて行う場合を含む。

【００２８】情報フィルタリング装置１は、新規ユーザからの文書が入力されると（ステップＳ１０１）、この文書中に出現する単語の種類を次元数とし、各単語の出現頻度に重み付けをしたベクトル（特徴ベクトル集合）を抽出する（ステップＳ１０２）。この抽出されたベクトルに対して上述の次元削減を行い（ステップＳ１０３）、これにより得られたベクトルから相関行列を作成する（ステップＳ１０４）。この相関行列は、所定の部分空間類別基準、すなわち、ベクトル空間モデルのパターン認識を行う場合に用いられる部分空間法に基づく基準に基づいて作成される。この部分空間類別基準については、例えば、「パターン認識と部分空間法」（エルッキ・オヤ著、産業図書）等の記載を参考にすることができる。この相関行列は、ユーザプロファイルとして保存

される（ステップS105）。このように、相関行列をユーザプロファイルとすることにより、ユーザの関心を表す関心情報は、文書中に出現する単語間の共起関係に着目した表現となる。

【0029】図3は、フィルタリングの処理手順図である。情報フィルタリング装置1は、文書が入力されると（ステップS201）、その文書に対して上記ステップS102～103と同様の処理を施し、次元削減されたベクトルを抽出する（ステップS202～S203）。そして、抽出したベクトルと、ユーザの識別情報を検索キーとしてプロファイルDB20から読み出したユーザプロファイルまたは共用プロファイルに対し、それぞれ前述のKL解析を施して固有値及び固有ベクトルを算出し、部分空間（プロファイル）に対するベクトルの射影を算出抽出することにより、文書の選別を行う（ステップS204）。

【0030】図4は、プロファイル更新の処理手順図である。この場合に入力される文書は、上記図3の処理手順により得られたフィルタリング結果である。このフィルタリング結果にも、「関心有」または「関心無」の識別情報が付与されている。フィルタリング結果である文書が入力されると（ステップS301）、情報フィルタリング装置1は、その文書に対して上記ステップS102～103と同様の処理を施し、次元削減を施したベクトルを抽出する（ステップS302～S303）。その後、抽出したベクトルからALSMの適応的な学習条件に基づいて相関行列を再作成する（ステップS304）。さらに、既に保存されている該当ユーザプロファイルに対応する相関行列と該当部分空間をこの再作成した相関行列で更新する（ステップS305）。このようにしてフィードバックの度にユーザプロファイルが自動的に更新される。

【0031】次に、共用プロファイルについてより詳しく説明する。共用プロファイルの作成、フィルタリング、更新も、基本的には上記図2～4の処理手順に従って行うことができる。ここでは、文書に付与された「関心有」と「関心無」の両面を考慮した共用プロファイルの作成を中心に説明する。なお、識別情報「関心有」に対応するカテゴリを「正の関心」、また「関心無」に対応するカテゴリを「負の関心」とする。関連する複数のユーザプロファイルの統合を行う際には、各ユーザプロファイル間における類似の度合い、すなわち距離値の大きさを考慮することが、共用プロファイルを効果的に作成する上で重要となる。

【0032】例えば、文書集合Dにおけるユーザプロファイルの、ユーザプロファイルBに対する距離値を抽出する場合に、「正の関心」及び「負の関心」に着目することにより、共用ファイルを高精度に作成することができる。この点を説明する。文書集合Dに文書Kが含まれている場合、まず、プロファイルBにおける文書Kに関して、「正の関心」の射影値から「負の関心」の射影値の差分を算出する（第1射影差分値）。次に、プロファイルAにおける文書Kに関して、「正の関心」の射影値から「負の関心」の射影値の差分を算出する（第2射影差分値）。さらに、プロファイルBの第1射影差分値からプロファイルAの第2射影差分値の差を算出する。この算出値を2乗した値を差分距離とする。この差分距離は、ユークリッド距離として表現される。文書Kに対するユーザプロファイルA及びBにおける関心の有無の差となるものである。

【0033】次に、同様にして、差分距離を文書集合Dに含まれるすべての文書に対して算出し、これらの各差分距離の総和を算出する。この総和による算出値が、文書集合Dにおけるユーザプロファイルのユーザプロファイルに対する距離値であり、関心の程度が似通った文書の順に足しあげる際の、すなわち統合する場合のユーザプロファイル間における差異の尺度となるものである。この距離値が大きい場合には対象となるユーザプロファイル間では類似度合いが小さく、一方、距離値が小さい場合にはユーザプロファイル間における類似度合いは大きくなる。共用ファイルを作成する際には、この類似度合いが所定範囲内のもの（例えば予め定めた閾値以下のもの）を統合するように構成する。

【0034】次に、情報フィルタリング装置1を用いた情報フィルタリング方法を、図5及び図6を参照して説明する。まず、初期プロファイルを作成する場合の例を説明する。文書がデータ入力部11を通じて入力され、新規ユーザであることを確認すると（ステップS401、S402）、ユーザ管理部12は、その文書に初期プロファイル設定用データを付与して特徴ベクトル抽出部13へ送る。特徴ベクトル抽出部13は、この文書からベクトル抽出を行う（ステップS403）。さらに、抽出したベクトルから次元処理部14で冗長な次元を削減する（ステップS404）。相関行列作成部16は、この次元削減されたベクトルから相関行列を作成し（ステップS406）、プロファイル管理部17に送る。プロファイル管理部17は、この相関行列をユーザプロファイル（初期プロファイル）として、プロファイルDB22に保存する（ステップS407）。

【0035】図6に移り、初期プロファイルの作成が終了した場合は、そのユーザが属すべき共用プロファイルを設定されているかどうかを調べる。共用プロファイルが設定されている場合は（ステップS408：Yes）、当該共用プロファイルが既に存在するか否かを調べ、存在しない場合には（ステップS409：No）、共用プロファイル処理部21で、対応する共用プロファイルを新規作成する（ステップS410）。当該共用プロファイルが既に存在する場合は（ステップS409：Yes）、その既存の共用プロファイルを、初期プロファイルの情報で更新する（ステップS411）。次の文書がある場

合はステップS401に戻り（ステップS412：Yes）、上記処理を繰り返す。初期プロファイルの作成だけを行う場合は処理を終了する（ステップS412：No）。

【0036】図5に戻り、この初期プロファイル、更新されて保存されているユーザプロファイル、あるいは共用プロファイル（便宜上、単にプロファイルとする）を用いてフィルタリングを行う場合は、選別対象となる文書に対してステップS402～S403の処理を施し、これにより得られたベクトルをフィルタリング処理部18に送る。フィルタリング処理部18は、このベクトルと該当するプロファイルとの射影を算出し（ステップS413）、算出結果に基づいて文書選別を行う（ステップS414）。選別結果は結果出力部19を通じてユーザに提示される。また、プロファイル更新のためにステップS401に戻る（ステップS415）。

【0037】プロファイル更新は、選別結果である文書に対してステップS402～S403の処理を施し、これにより得られたベクトルを相関行列作成部16に送る。相関行列作成部16は、このベクトルに基づいて相関行列を再作成する（ステップS416）。プロファイル管理部17は、対応するプロファイルを再作成した相関行列で更新する（ステップS417）。共用プロファイルに関する処理（図6参照）は、初期プロファイルの作成の場合と同様となる。

【0038】なお、本実施形態では、相関行列をユーザプロファイルとしているが、これは入力文書における単語の共起関係に着目し、共用プロファイルとして各ユーザプロファイルの合成の際の処理負荷を軽減することを目的とするものである。しかしながらこの手法以外にも、例えば、相関行列に対してKL解析を施し、その結果得られるベクトル空間をユーザプロファイとして作成することも可能である。この場合の情報量は、ユーザプロファイルに相関行列を用いる場合とほぼ同じになるが、共用プロファイルとして各ユーザプロファイルを合成する際に、次元数を統一するための処理が必要となるものである。

【0039】（第2実施形態）本発明は、通信回線としてインタネット等の公衆網を介して流通する大量の電子化情報に対して自動的なフィルタリングを行うシステム、例えば、上記情報フィルタリング装置として機能する情報フィルタリングサーバ、公衆網から情報を取得する機能を有するクライアントを配備した情報フィルタリングシステムの形態で実施することも可能である。

【0040】この場合の情報フィルタリングサーバは、例えば、インタネット環境上における複数の大規模なデータベースを具備した各種情報提供サーバからの電子化情報からクライアントに最適な情報を選択して提供する情報提供支援サーバ、所謂、情報ナビゲーションサーバとして位置付けることができる。

【0041】この場合の構成例としては、コンピュータ装置の内部あるいは外部記憶装置に、上記プロファイルDB20と同一のデータベースを構築し、公衆網を介してクライアント及び上記各種情報提供サーバとの通信を行う通信制御部を具備する。さらに上記情報フィルタリング装置1と同様の機能ブロック、すなわち、データ入力部11、ユーザ管理部12、特徴ベクトル抽出部13、次元処理部14、処理選択部15、相関行列作成部16、プロファイル管理部17、フィルタリング処理部18、結果出力部19、を具備して構成する。

【0042】この情報フィルタリングサーバが上記情報フィルタリング装置1と相違する点は、通信制御を行う公知の通信制御部を具備する点である。この通信制御部を介して流通する電子化情報群をデータ入力部11に入力し、クライアントからの情報取得要求を受け付けるように構成することで、ネットワークを用いた情報フィルタリングが可能になる。この場合の情報取得要求の入力は、例えばWWW環境のブラウザ等をインタフェースとして使用することができる。また、上記各種情報提供サーバからの電子化情報群は、必ず情報フィルタリングサーバを経由するようにし、この電子化情報群に対する選別結果を通信制御部を介してクライアントに提供するように構成する。また、情報フィルタリングサーバは、例えば、インタネット環境におけるサーバのエージェント技術と融合することにより、流通する大量の電子化情報群に対して自動的なフィルタリングを行うシステムの構築が可能になる。

【0043】次に、実際に、本実施形態の情報フィルタリング装置1における評価実験を行った結果について説明する。この実験では、評価用文書として1995年1月から1996年10月までの1年間分の日本語による新聞記事を使用し、前半年分を初期プロファイル用設定データの文書集合（以下、訓練集合）、また後半年分を評価実験用の文書集合（以下、評価集合）とした。

【0044】これらの記事は、図7に示す5つのジャンルのいずれかに属するとともに、図中の記事数が各ジャンル毎のデータ数を表している。また各記事は、タイトルと本文から構成され、当該記事が属するジャンル名が付与されているものとする。本実験では、単語を特定し、訓練集合においてその単語を含むものを「関心有」、その他を「関心無」の記事と便宜的に分けてユーザプロファイルを生成した。ここでは、これらの単語を関心事項として「トピック」とし、対応するトピックによるプロファイルで評価集合をフィルタリングした実験結果を示すものである。

【0045】図8は、単語「輸入」をトピックにフィルタリングして得られた1225記事のうち、25記事のタイトルを示している。この結果によれば、トピック「輸入」と意味の類似する単語「貿易」に関する記事のほか、必ずしも語彙上関連深くない単語「病気」に関す

る記事が比較的目立つことがわかる。これは、「輸入」の語を含む記事中に同時に「感染」の語を含む記事が多いために「感染」も併せて関心事項とみなされたことによるものと解釈される。このことから、本実施形態の情報フィルタリング装置１によれば、ユーザが関心事項の内容や数を明確に意識しなくても、学習した記事集合における語のつながりを通して自動的に目的となる記事が抽出されることがわかる。

【００４６】図９及び図１０は、単語「阪神」をトピックにフィルタリングされた記事中において、「プロ野球」及び「地震」の単語を含む記事の分布を、「正の関心」の部分空間における第３成分までの値で表したものである。この場合、「プロ野球」及び「地震」の単語を含む記事数は、「阪神」をトピックにフィルタリングされた４６記事中で、各々、"１０６４"及び"１２３"である。また、図中における各点は、１記事毎に対応しており、各成分へのマッピング結果を表している。

【００４７】図９の分布結果では、第１成分と第２成分に広がりをもち、特に第３成分に偏った散らばりがみられる。一方、図１０の分布結果では、第３成分はすべてゼロに近い値をとり、第１及び第２成分の平面で大きく広がる対照的な分布となっている。従って、記事中における単語のつながりと部分空間の軸に対応がとられ、このことが、上述のような複数の関心事項のフィルタリングを可能にしていることがわかる。

【００４８】このように、本実施形態の情報フィルタリング装置１では、文書中における単語の共起関係が導入され、関心有り及び無しの両面を考慮した類別基準によるフィルタリングを行うことにより、従来型のような一面的な情報フィルタリングと比較して、より精度の高い選別結果を得ることが可能となる。

【００４９】また、作成される相関行列をユーザプロファイルとして用いることから、ユーザの関心事項の数や内容に限定されることなく、それらの関連に応じたフィルタリングが単一のユーザプロファイルにより得られるようになり、また、ユーザの関心事項の広がりを部分空間における次元数から知ることができるようになる。

【００５０】また、相関行列に対する適応的な学習により、ユーザプロファイルの表現を変えることなくユーザの関心事項の変化に対して柔軟に追随できるので、関心の時間的な変化に対応した自動的なフィルタリングが可能となる。

【００５１】また、共用プロファイルを各ユーザプロファイルにおける相関行列の和から生成し、ユーザ間における関心事項の共有化を図ることが可能なので、ユーザ個人に特定して絞り込んだ関心情報に基づいたために取りこぼした情報や、ユーザに近い関心情報でありながら、当該ユーザの関心情報からは抽出不可能であった情報に対する補完的なフィルタリングが可能となる。

【００５２】また、個々の各ユーザプロファイルの情報から、サービス提供者等のシステム運用管理者は、ユーザ全体における関心の動向が把握可能となり、対象となるユーザに応じた、例えば、広告やダイレクトメール等のダイレクトマーケッティングや、新商品開発のマーケッティングリサーチ用の調査材料となり得る効果がある。

【００５３】さらに、既存の複数の情報提供サービスシステム等と独立して動作するシステムの構築や、既存システムへの組み込みも容易になる。

【００５４】

【発明の効果】以上の説明から明らかなように、本発明によれば、ユーザの関心情報であるユーザプロファイルが自動的に作成される効果がある。また、情報フィルタリングに際して、ユーザの関心を「関心有」と「関心無」の両面が考慮されるので、フィルタリングの精度を一定値以上に維持することが可能となる。

【００５５】本発明をネットワーク環境下で適用させた場合には、この情報フィルタリングにより、継続的に流入する大量の電子化情報群からユーザの関心に基づいて確実且つタイムリーに必要となる情報の取得が出来ることから、情報の有効活用が促進される。このことから、アクセス効率及び実用性が格段に向上するシステムの提供が可能となる。

【図面の簡単な説明】

【図１】本発明の一実施形態に係る情報フィルタリング装置の機能ブロック図。

【図２】プロファイル作成の処理手順図。

【図３】フィルタリングの処理手順図。

【図４】プロファイル更新時の処理手順図。

【図５】本発明の情報フィルタリング装置によるフィルタリング方法の手順説明図。

【図６】本発明の情報フィルタリング装置によるフィルタリング方法の手順説明図。

【図７】本発明の情報フィルタリング装置の評価実験データを示した図表。

【図８】「輸入」のトピックでフィルタリングされた評価集合のタイトル群。

【図９】「阪神」のトピックでフィルタリングされた「プロ野球」の語を含む記事の分布結果。

【図１０】「阪神」のトピックでフィルタリングされた「地震」の語を含む記事の分布結果。

【符号の説明】

１　情報フィルタリング装置
１１　データ入力部
１２　ユーザ管理部
１３　特徴ベクトル抽出部
１４　次元処理部
１５　処理選択部
１６　相関行列作成部
１７　プロファイル管理部

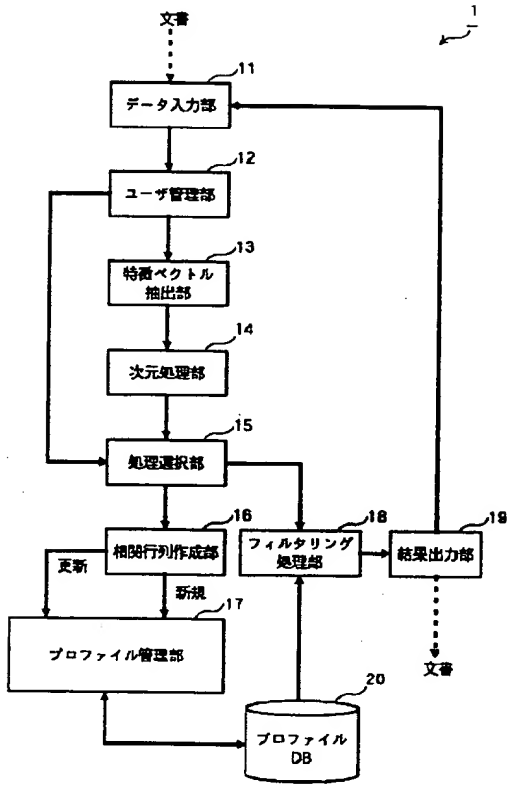
15

１８　フィルタリング処理部
１９　結果出力部

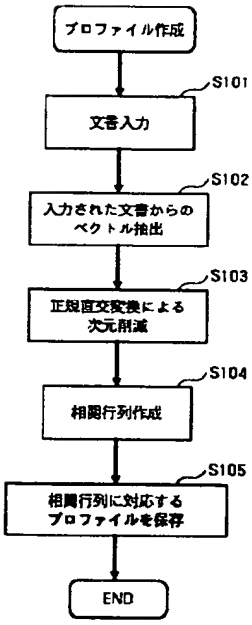
15

＊２０　プロファイルＤＢ（データベース）
＊

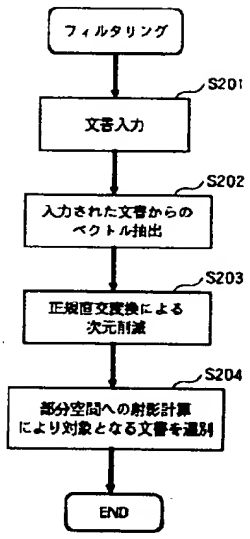
【図１】



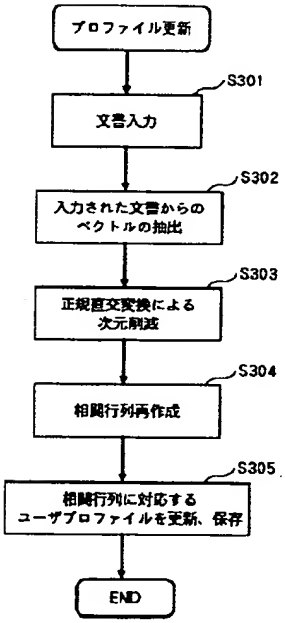
【図２】



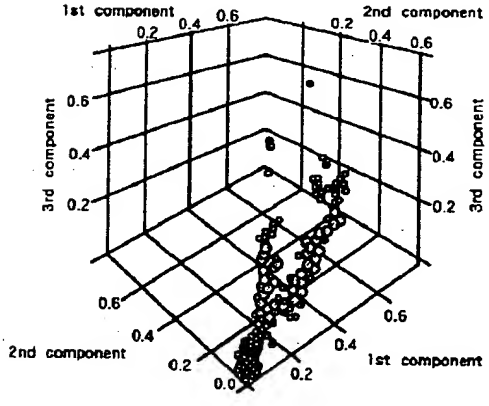
【図３】



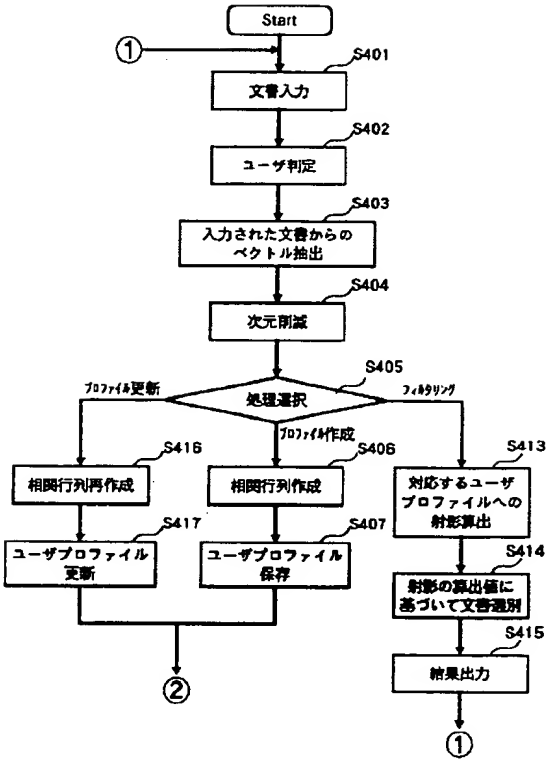
【図４】



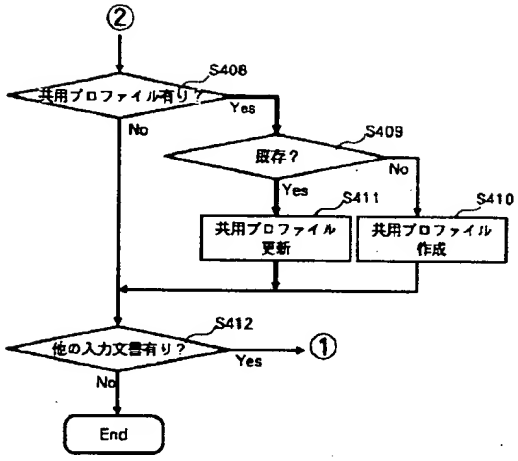
【図９】



【図5】



【図6】



【図7】

記事数	スポーツ	経済	国際	社会	政治	総計
訓練集合	2736	2639	2397	5453	2638	15803
評価集合	3794	3213	3493	5880	3131	19691

【図8】

1　［リコール］　＊国＊＊＊社製４車種届け出
2　［リコール］　＊＊＊＊製２車種に欠陥
3　［欠陥車］　＊＊＊７車種２５万台リコール
4　［リコール］　＊＊＊＊が３車種届け出
5　［リコール］　＊＊＊＊軽貨物車の２車種
6　［薬害エイズ］　＊＊元課長が８４年に非加熱製剤の危険性を指摘
7　［自主回収］　＊国で細菌混入の製剤を－－＊＊＊製薬等
8　［異物混入］　＊＊＊の清涼飲料水４８万本回収へ
9　［貿易統計］　黒字が６期連続で前年同期下回る
10　［薬害エイズ］　血友病以外の非加熱製剤投与、集合鐘ぐるみ－＊＊省
11　［薬害エイズ］　非加熱製剤、血友病以外の患者に広範囲で使われる
12　［薬害エイズ］　＊＊省対応に疑問－－＊厚相
13　［血液製剤］　ヤコブ病感染の疑いで２万本回収－－＊＊＊＊－
14　［薬害エイズ］　汚吏との戦い振り返る－－＊＊医学会総会で－＊＊相
15　［薬害エイズ］　薬害エイズ＝ことば説明
16　［薬害エイズ］　血友病専門医らにも大きな衝撃
17　［７月の貿易統計］　黒字額２０カ月連続の減
18　［エイズ禍］　２カ月に患者・感染者１１２人過去最多
19　［Ｏ１５７］　集団感染とＤＮＡパターン一致へ
20　［薬害エイズ］　＊＊＊副社長が昇格へ
21　［自動車産業］　＊＊２０車種火災の恐れ
22　［院内感染］　世界的に希な事態に関係者ら衝撃－－＊＊大学病院
23　［貿易統計］　黒字、前年同期比４１．６％減と７期連続減少
24　［特報・薬害エイズ］　非血友病への投与
25　［狂牛病］　同症状の羊の病気

【図10】